Clin Orthop Relat Res (2010) 468:885–892
DOI 10.1007/s11999-009-1164-4

CLINICAL RESEARCH

P Value and the Theory of Hypothesis Testing

An Explanation for New Researchers

David Jean Biau MD, Brigitte M. Jolles MD Msc, MD, Raphaël Porcher PhD

Received: 11 February 2009 / Accepted: 2 November 2009 / Published online: 17 November 2009

Abstract In the 1920s, Ronald Fisher developed the theory behind the p value and Jerzy Neyman and Egon Pearson developed the theory of hypothesis testing. These distinct theories have provided researchers important quantitative tools to confirm or refute their hypotheses. The p value is the probability to obtain an effect equal to or more extreme than the one observed presuming the null hypothesis of no effect is true; it gives researchers a measure of the strength of evidence against the null hypothesis. As commonly used, investigators will select a threshold p value below which they will reject the null hypothesis. The theory of hypothesis testing allows researchers to reject a null hypothesis in favor of an alternative hypothesis of some effect. As commonly used, investigators choose Type I error (rejecting the null hypothesis when it is true) and Type II error (accepting the null hypothesis when it is false) levels and determine some critical region. If the test statistic falls into that critical region, the null hypothesis is rejected in favor of the alternative hypothesis. Despite similarities between the two, the p value and the theory of hypothesis testing are different theories that often are misunderstood and confused, leading researchers to improper conclusions. Perhaps the most common misconception is to consider the p value as the probability that the null hypothesis is true rather than the probability of obtaining the difference observed, or one that is more extreme, considering the null is true. Another concern is the risk that an important proportion of statistically significant results are falsely significant. Researchers should have a minimum understanding of these two theories so that they are better able to plan, conduct, interpret, and report scientific experiments.

Each author certifies that he or she has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

D. J. Biau (✉), R. Porcher
Département de Biostatistique et Informatique Médicale, INSERM–UMR-S 717, AP-HP, Université Paris 7, Hôpital Saint Louis, 1, Avenue Claude-Vellefaux, Paris Cedex 10 75475, France
e-mail: djmbiau@yahoo.fr

B. M. Jolles
Hôpital Orthopédique Département de l'Appareil Locomoteur Centre Hospitalier, Universitaire Vaudois Université de Lausanne, Lausanne, Switzerland

Introduction

> *"We are inclined to think that as far as a particular hypothesis is concerned, no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis"* [15].

Since their introduction in the 1920s, the p value and the theory of hypothesis testing have permeated the scientific community and medical research almost completely. These theories allow a researcher to address a certain hypothesis such as the superiority of one treatment over another or the association between a characteristic and an outcome. In these cases, researchers frequently wish to disprove the well-known null hypothesis, that is, the absence of difference between treatments or the absence of association of a characteristic with outcome. Although statistically the null hypothesis does not necessarily relate to no effect or to no association, the presumption that it does relate to no effect or association frequently is made in medical research and

the one we will consider here. The introduction of these theories in scientific reasoning has provided important quantitative tools for researchers to plan studies, report findings, compare results, and even make decisions. However, there is increasing concern that these tools are not properly used [9, 10, 13, 20].

The p value is attributed to Ronald Fisher and represents the probability of obtaining an effect equal to or more extreme than the one observed considering the null hypothesis is true [3]. The lower the p value, the more unlikely the null hypothesis is, and at some point of low probability, the null hypothesis is preferably rejected. The p value thus provides a quantitative strength of evidence against the null hypothesis stated.

The theory of hypothesis testing formulated by Jerzy Neyman and Egon Pearson [15] was that regardless of the results of an experiment, one could never be absolutely certain whether a particular treatment was superior to another. However, they proposed one could limit the risks of concluding a difference when there is none (Type I error) or concluding there is no difference when there is one (Type II error) over numerous experiments to prespecified chosen levels denoted α and β, respectively. The theory of hypothesis testing offers a rule of behavior that, in the long run, ensures followers they would not be wrong often.

Despite simple formulations, both theories frequently are misunderstood and misconceptions have emerged in the scientific community. Therefore, researchers should have a minimum understanding of the p value and hypothesis testing to manipulate these tools adequately and avoid misinterpretation and errors in judgment. In this article, we present the basic statistics behind the p value and hypothesis testing, with historical perspectives, common misunderstandings, and examples of use for each theory. Finally, we discuss the implications of these issues for clinical research.

The p Value

The p value is the probability of obtaining an effect equal to or more extreme than the one observed considering the null hypothesis is true. This effect can be a difference in a measurement between two groups or any measure of association between two variables. Although the p value was introduced by Karl Pearson in 1900 with his chi square test [17], it was the Englishman Sir Ronald A. Fisher, considered by many as the father of modern statistics, who in 1925 first gave the means to calculate the p value in a wide variety of situations [3].

Fisher's theory may be presented as follows. Let us consider some hypothesis, namely the null hypothesis, of no association between a characteristic and an outcome. For any magnitude of the association observed after an experiment is conducted, we can compute a test statistic that measures the difference between what is observed and the null hypothesis. This test statistic may be converted to a probability, namely the p value, using the probability distribution of the test statistic under the null hypothesis. For instance, depending on the situation, the test statistic may follow a χ^2 distribution (chi square test statistic) or a Student's t distribution. Its graphically famous form is the bell-shaped curve of the probability distribution function of a t test statistic (Fig. 1A). The null hypothesis is said to be disproven if the effect observed is so important, and consequently the p value is so low, that "either an exceptionally rare chance has occurred or the theory is not true" [6]. Fisher, who was an applied researcher, strongly believed the p value was solely an objective aid to assess the plausibility of a hypothesis and ultimately the conclusion of differences or associations to be drawn remained to the scientist who had all the available facts at hand. Although he supported a p value of 0.05 or less as indicating evidence against the null, he also considered other more stringent cutoffs. In his words "If p is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05…" [4].

For instance, say a researcher wants to test the association between the existence of a radiolucent line in Zone 1 on the postoperative radiograph in cemented cups and the risk of acetabular loosening. He or she can use a score test in a Cox regression model, after adjusting for other potentially important confounding variables. The null hypothesis that he or she implicitly wants to disprove is that a radiolucent line in Zone 1 has no effect on acetabular loosening. The researcher's hypothetical study shows an increased occurrence of acetabular loosening when a radiolucent line in Zone 1 exists on the postoperative radiograph and the p value computed using the score test is 0.02. Consequently, the researcher concludes either a rare event has occurred or the null hypothesis of no association is not true. Similarly, the p value may be used to test the null hypothesis of no difference between two or more treatments. The lower the p value, the more likely is the difference between treatments.

The Neyman-Pearson Theory of Hypothesis Testing

We owe the theory of hypothesis testing as we use it today to the Polish mathematician Jerzy Neyman and American statistician Egon Pearson (the son of Karl Pearson). Neyman and Pearson [15] thought one could not consider a null hypothesis unless one could conceive at least one plausible alternative hypothesis.

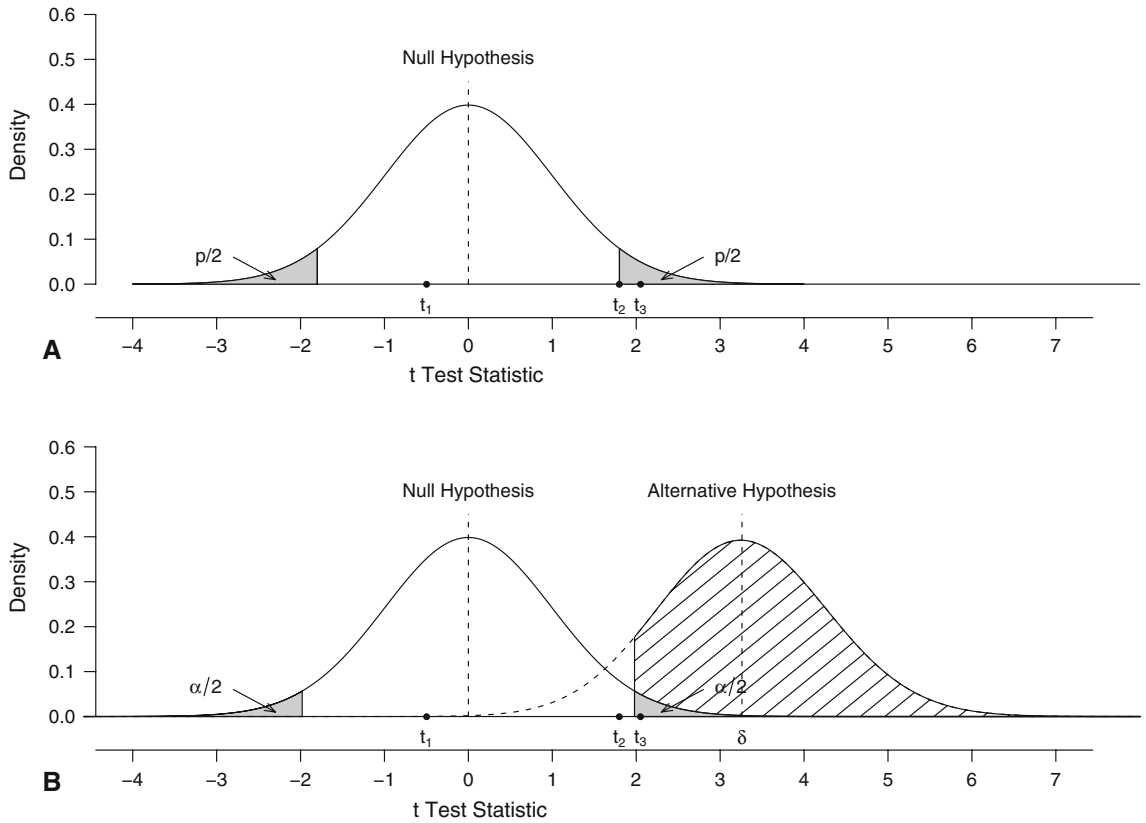

Fig. 1A–B These graphs show the results of three trials (t_1, t_2, and t_3) comparing the 1-month HHS after miniincision or standard incision hip arthroplasty under the theory of (**A**) Fisher and (**B**) Neyman and Pearson. For these trials, $\alpha = 5\%$ and $\beta = 10\%$. Trial 1 yields a standardized difference between the groups of 0.5 in favor of the standard incision; Trials 2 and 3 yield standardized differences of 1.8 and 2.05, respectively. The corresponding p values are 0.62, 0.074, and 0.042 for Trials 1, 2, and 3, respectively. (**A**) Fisher's p value for Trial 2 is represented by the gray area under the null hypothesis; it corresponds to the probability of observing a standardized difference of 1.8 (Point 2) or more extreme differences (gray area on both sides) considering the null hypothesis is true. According to Fisher, Trials 2 and 3 provide fair evidence against the null hypothesis of no difference between treatments; the decision to reject the null hypothesis of no difference in these cases will depend on other important information (previous data, etc). Trial 1 provides poor evidence against the null as the difference observed, or one more extreme, had 62% probability of resulting from chance alone if the treatments were equal. (**B**) Under the Neyman and Pearson theory, the Types I ($\alpha = 0.05$, gray area under the null hypothesis) and II ($\beta = 0.1$, shaded area under the alternative hypothesis) error rates and the difference to be detected ($\delta = 10$) define a critical region for the test statistic ($|t\ test| > 1.97$). If the test statistic (standardized difference here) falls into that critical region, the null hypothesis is rejected; this is the case for Trial 3. Trials 1 and 2 do not fall into the critical region and the null is not rejected. According to Neyman and Pearson's theory, the null hypothesis of no difference between treatments is rejected after Trial 3 only. The distributions depicted are the probability distribution functions of the t test with 168 degrees of freedom.

Their theory may be presented in a few words this way. Consider a null hypothesis H0 of equal improvement for patients under Treatment A or B and an alternative hypothesis H1 of a difference in improvement of some relevant size δ between the two treatments. Researchers may make two types of incorrect decisions at the end of a trial: they may consider the null hypothesis false when it is true (a Type I error) or consider the null true when it is in fact false (Type II error) (Table 1). Neyman and Pearson proposed, if we set the risks we are willing to accept for Type I errors, say α (ie, the probability of a Type I error), and Type II errors, say β (ie, the probability of a Type II error), then, "without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong." These Types I and II error rates

Table 1. Types I and II errors according to the theory of hypothesis tests

Study findings	Truth	
	Null hypothesis is true	Null hypothesis is false
Null hypothesis is not rejected	True negative	Type II error (β) (false negative)
Null hypothesis is rejected	Type I error (α) (false positive)	True positive

α and β represent the probability of Types I and II errors, respectively.

allow defining a critical region for the test statistic used. For instance for α set at 5%, the corresponding critical regions would be $\chi^2 > 3.84$ for the chi square statistic or $|t_{168df}| > 1.97$ for Student's t test with 168 degrees of freedom (Fig. 1B) (the reader need not know the details of these computations to grasp the point). If, for example, the comparison of the mean improvement under Treatments A and B falls into that critical region, then the null hypothesis is rejected in favor of the alternative; otherwise, the null hypothesis is accepted. In the case of group comparisons, the test statistic represents a measure of the likelihood that the groups compared are issued from the same population (null hypothesis): the more groups differ, the higher the test statistic and at some point the null hypothesis is rejected and the alternative is accepted. Although Neyman and Pearson did not view the 5% level for Type I error as a binding threshold, this level has permeated the scientific community. For the Type II error rate, 0.1 or 0.2 often is chosen and corresponds to powers (defined as $1 - \beta$) of 90% and 80%, respectively.

For instance, say a surgeon wants to compare the 1-month Harris hip score (HHS) after miniincision and standard incision hip arthroplasty. With the help of a statistician, he plans a randomized controlled trial and considers the null hypothesis H0 of no difference between the standard treatment and experimental treatment (miniincision) and the alternative hypothesis H1 of a difference δ of more than 10 points on the HHS, which he considers is the minimal clinically important difference. Because the statistician is performing many statistical tests across different studies all day long, she has grown very concerned about false positives and, as a general rule, she is not willing to accept more than 5% Type I error rate, that is, if no difference exists between treatments, there is only a 5% chance to conclude a difference. However, the surgeon is willing to give the best chances to detect that difference if it exists and chooses a Type II error of 10%, ie, a power of 90%; therefore, if a difference of 10 points exists between treatments, there is an acceptable 10% chance that the trial will not detect it. Let us presume the expected 1-month HHS after standard incision hip arthroplasty is 70 and the expected SD in both groups is 20. The required sample size therefore is 85 patients per group (two-sample t test).The critical region to reject the null hypothesis therefore is 1.97 (Student's t test with 168 degrees of freedom). Therefore, if at the end of the trial Student's t test yields a statistic of 1.97 or greater, the null hypothesis will be rejected; otherwise the null hypothesis will not be rejected and the trial will conclude no difference between the experimental and standard treatment groups. Although the Neyman-Pearson theory of hypothesis testing usually is used for group comparisons, it also may be used for other purposes such as to test the association of a variable and an outcome.

The Difference between Fisher's P Value and Neyman-Pearson's Hypothesis Testing

Despite the fiery opposition these two schools of thought have concentrated against each other for more than 70 years, the two approaches nowadays are embedded in a single exercise that often leads to misuse of the original approaches by naïve researchers and sometimes even statisticians (Table 2) [13]. Fisher's significance testing with the p value is a practical approach whose statistical properties are derived from a hypothetical infinite population and which applies to any single experiment. Neyman and Pearson's theory of hypothesis testing is a more mathematical view with statistical properties derived from the long-run frequency of experiments and does not provide by itself evidence of the truth or falsehood of a particular hypothesis. The confusion between approaches comes from the fact that the critical region of Neyman-Pearson theory can be defined in terms of p value. For instance, the critical regions defined by thresholds at $\pm$ 1.96 for the normal distribution, 3.84 for the chi square test at 1 degree of freedom, and $\pm$ 1.97 for a t test at 168 degrees of freedom all correspond to setting a threshold at 0.05 for the p value. The p value is found more practical because it represents a single probability across the different distributions of numerous test statistics and usually the value of the test statistic is omitted and only the p value is reported.

The difference between approaches may be more easily understandable through a hypothetical example. After a trial comparing an experimental Treatment A with a standard Treatment B is conducted, a surgeon has to decide whether Treatment A is or is not superior to Treatment B. Following Fisher's theory, the surgeon weighs issues such

Table 2. Comparison of Fisher's p value and Neyman-Pearson's hypothesis testing

Fisher's p value	Hypothesis testing
Ronald Fisher	Jerzy Neyman and Egon Pearson
Significance test	Hypothesis test
p Value	α
The p value is a measure of the evidence against the null hypothesis	α and β levels provide rules to limit the proportion of errors
Computed a posteriori from the data observed	Determined a priori at some specified level
Applies to any single experiment	Applies in the long run through the repetition of experiments
Subjective decision	Objective behavior
Evidential, ie, based on the evidence observed	Nonevidential, ie, based on a rule of behavior

as relevant in vitro tests, the design of the trial, previous results comparing treatments, etc, and the p value of the comparison to eventually reach a conclusion. In such cases, p values of 0.052 and 0.047 likely would be similarly weighted in making the conclusion whereas p values of 0.047 and 0.0001 probably would have differing weights. In contrast, statisticians have to give their opinion regarding an enormous quantity of new drugs and medical devices during their life. They cannot be concerned whether each new particular treatment tested is superior to the standard one because they know the evidence can never be certain. However, they know following Neyman and Pearson's theory they can control the overall proportion of errors, either Type I or II errors (Table 1), they make over their entire career. By setting α at, say, 5% and power $(1 - \beta)$ at 90%, at the end of their career, they know in 5% cases they will have concluded the experimental treatment was superior to the standard when it was not and in 10% cases they will have concluded the experimental treatment was not different from the standard treatment although it was. In that case, very close p values such as 0.047 and 0.052, will lead to rather dramatically opposite actions. In the first case, the treatment studied will be considered superior and used, when in the second case the treatment will be rejected for inefficacy despite very close evidence observed from the two experiments (in a Fisherian point of view).

Misconceptions When Considering Statistical Results

First, the most common and certainly most serious error made is to consider the p value as the probability that the null hypothesis is true. For instance, in the above-mentioned example to illustrate Fisher's theory, which yielded a p value of 0.02, one should not conclude the data show there is a 2% chance of no association between the existence of a radiolucent line in Zone 1 on the postoperative radiograph in cemented cups and the risk of acetabular loosening. The p value is not the probability of the null hypothesis being true; it is the probability of observing these data, or more extreme data, if the null is true. The p value is computed on the basis that the null hypothesis is true and therefore it cannot give any probability of it being more or less true. The proper interpretation in the example should be: considering no association exists between a radiolucent line in Zone 1 and the risk of acetabular loosening (the null hypothesis), there was only a 2% chance to observe the results of the study (or more extreme results).

Second, there is also a false impression that if trials are conducted with a controlled Type I error, say 5%, and adequate power, say 80%, then significant results almost always are corresponding to a true difference between the treatments compared. This is not the case, however. Imagine we test 1000 null hypotheses of no difference between experimental and control treatments. There is some evidence that the null only rarely is false, namely that only rarely the treatment under study is effective (either superior to a placebo or to the usual treatment) or that a factor under observation has some prognostic value [12, 19, 20]. Say that 10% of these 1000 null hypotheses are false and 90% are true [20]. Now if we conduct the tests at the aforementioned levels of $\alpha = 5\%$ and power = 80%, 36% of significant p values will not report true differences between treatments (Fig. 2, Scenario 1, 64% true-positive and 36% false-positive significant results; Fig. 3, Point A). Moreover, in certain contexts, the power of most studies does not exceed 50% [1, 7]; in that case, almost ½ of significant p values would not report true differences [20] (Fig. 3, Point B).

Implications for Research

Fisher, who designed studies for agricultural field experiments, insisted "a scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance" [5]. There are three issues that a researcher should consider when conducting, or when assessing the report of, a study (Table 3).

First, the relevance of the hypothesis tested is paramount to the solidity of the conclusion inferred. The proportion of false null hypotheses tested has a strong effect on the predictive value of significant results. For instance, say we shift from a presumed 10% of null hypotheses tested being false to a reasonable 33% (ie, from 10% of treatments tested effective to ⅓ of treatments tested effective), then the positive predictive value of significant results improves from 64% to 89% (Fig. 3, Point C). Just as a building cannot be expected to have more resistance to environmental challenges than its own foundation, a study nonetheless will fail regardless of its design, materials, and statistical analysis if the hypothesis tested is not sound. The danger of testing irrelevant or trivial hypotheses is that, owing to chance only, a small proportion of them eventually will wrongly reject the null and lead to the conclusion that Treatment A is superior to Treatment B or that a variable is associated with an outcome when it is not. Given that positive results are more likely to be reported than negative ones, a misleading impression may arise from the literature that a given treatment is effective when it is not and it may take numerous studies and a long time to invalidate this incorrect evidence. The requirement to register trials before the first patient is included may prove

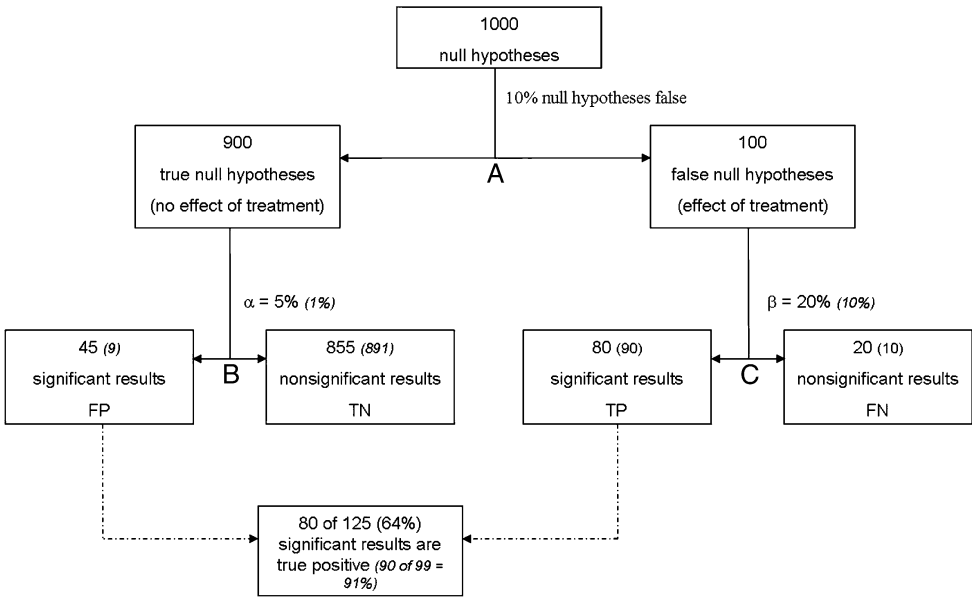

Fig. 2 The flowchart shows the classification tree for 1000 theoretical null hypotheses with two different scenarios considering 10% false null hypotheses. Scenario 1 has a Type I error rate of 5% and a Type II error rate of 20% (power = 80%); Scenario 2 has a Type I error rate of 1% and a Type II error rate of 10% (power = 90%). The first node (A) separates the 900 true null hypotheses (no effect of treatment) from the 100 false null hypotheses (effect of treatment). For Scenario 1, the second node left (B) separates the 900 true null hypotheses (no treatment effect) at the 5% level: 855 tests are not significant (true-negative [TN] results) and 45 tests are falsely significant (false-positive [FP] results). The second node right (C) separates the 100 false null hypotheses (effect of treatment) at the 20% level (power = 80%): 20 tests are falsely not significant (false-negative [FN] results) and 80 tests are significant (true-positive [TP] results). The corresponding positive predictive value [TP/(TP + FP)] is 64%. The figures in parentheses at the second nodes right and left and at the bottom show the results for Scenario 2. The positive predictive value of significant results for Scenario 2 is 91%.

to be an important means to deter this issue. For instance, by 1981, 246 factors had been reported [12] as potentially predictive of cardiovascular disease, with many having little or no relevance at all, such as certain fingerprints patterns, slow beard growth, decreased sense of enjoyment, garlic consumption, etc. More than 25 years later, only the following few are considered clinically relevant in assessing individual risk: age, gender, smoking status, systolic blood pressure, ratio of total cholesterol to high-density lipoprotein, body mass index, family history of coronary heart disease in first-degree relatives younger than 60 years, area measure of deprivation, and existing treatment with antihypertensive agent [19]. Therefore it is of prime importance that researchers provide the a priori scientific background for testing a hypothesis at the time of planning the study, and when reporting the findings, so that peers may adequately assess the relevance of the research. For instance, with respect to the first example given, we may hypothesize that the presence of a radiolucent line observed in Zone 1 on the postoperative radiograph is a sign of a gap between cement and bone that will favor micromotion and facilitate the passage of polyethylene wear particles, both of which will favor eventual bone resorption and loosening [16, 18]. An important endorsement exists when other studies also have reported the association [8, 11, 14].

Second, it is essential to plan and conduct studies that limit the biases so that the outcome rightfully may be attributed to the effect under observation of the study. The difference observed at the end of an experiment between two treatments is the sum of the effect of chance, of the treatment or characteristic studied, and of other confounding factors or biases. Therefore, it is essential to minimize the effect of confounding factors by adequately planning and conducting the study so we know the difference observed can be inferred to be the treatment studied, considering we are willing to reject the effect of chance (when the p value or equivalently the test statistic engages us to do so). Randomization, when adequate, for example, when comparing the 1-month HHS after miniincision and standard incision hip arthroplasty, is the preferred experimental design to control on average known and unknown confounding factors. The same principles should apply to other experimental designs. For instance, owing to the rare and late occurrence of certain events, a retrospective study rather than a prospective study is preferable to judge the association between the existence of a radiolucent line in Zone 1 on the postoperative radiograph in cemented cups and the risk of acetabular loosening. Nonetheless researchers should ensure there is no systematic difference regarding all known confounding factors between patients who have a radiolucent line in

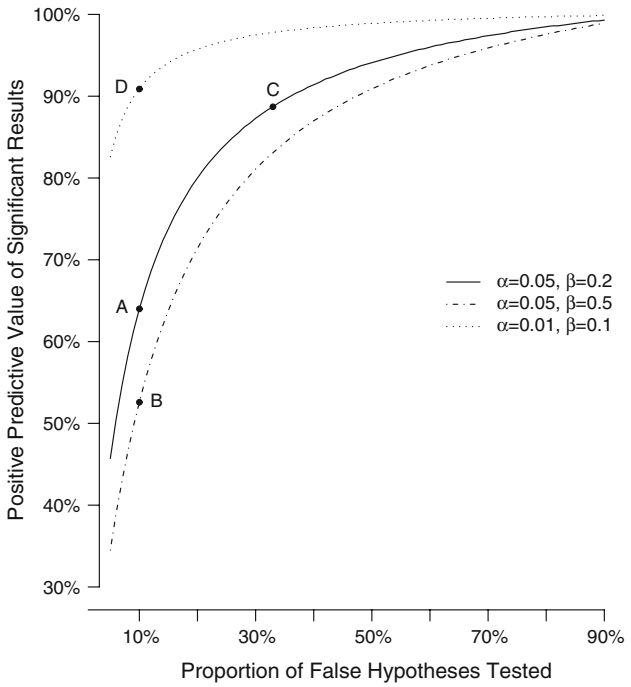

Fig. 3 This graph shows the effect of the Types I and II error rates and the proportion of false null hypotheses (true effect of treatment) on the positive predictive value of significant results. Three different levels of Types I and II error rates are depicted: $\alpha = 5\%$ and $\beta = 20\%$ (power = 80%), $\alpha = 5\%$ and $\beta = 50\%$ (power = 50%), and $\alpha = 1\%$ and $\beta = 10\%$ (power = 90%). It can be seen, the higher the proportion of false null hypotheses tested, the better is the positive predictive value of significant results. Point A corresponds to a standard $\alpha = 5\%$, $\beta = 20\%$ (power = 80%), and 10% of false null hypotheses tested. The positive predictive value of a significant result is 64% (also see Fig. 2). Point B corresponds to the suspected reality $\alpha = 5\%$, $\beta = 50\%$ (power = 50%), and 10% of false null hypotheses tested. The positive predictive value of a significant result decreases to 53%. Point C corresponds to $\alpha = 5\%$, $\beta = 20\%$ (power = 80%), and 33% of false null hypotheses tested. The positive predictive value of a significant result increases to 89%. Finally, Point D corresponds to $\alpha = 1\%$, $\beta = 10\%$ (power = 90%), and 10% of false null hypotheses tested. The positive predictive value of a significant result increases to 91%. At the extreme, if all null hypotheses tested are true (no effect of treatment), regardless of α and β, the positive predictive value of a significant result is 0.

Zone 1 and those who do not. For instance, they should retrieve both groups over the same period of time and the acetabular components used and patients under study should be the same in both groups. If the types of acetabular components differ markedly between groups, the researcher will not be able to say whether the difference observed in aseptic loosening between groups is attributable to the existence of a radiolucent line in Zone 1 or to differences in design between acetabular components.

Last, choosing adequate levels of Type I and Type II errors, or alternatively the level of significance for the p value, may improve the reliance we can have in purported significant results (Figs. 2, 3). Decreasing the α level will proportionally decrease the number of false-positive findings and subsequently improve the positive predictive value of significant results. Increasing the power of studies will correspondingly increase the number of true-positive findings and also improve the positive predictive value of significant results. For example, if we shift from a Type I error rate of 5% and power of 80% to a Type I error rate of 1% and power of 90%, the positive predictive value of a significant result increases from 64% to 91% (Fig. 2, Scenario 2; Fig. 3, Point D). Sample size can be used as a lever to control for Types I and II error levels [2]. However, a strong statistical association, p values, or test statistics never imply any causal effect. The causal effect is built on, study after study, little by little. Therefore, replication of the experiment by others is crucial before accepting any hypothesis. To replicate an experiment, the methods used must be described sufficiently so that the study can be replicated by other informed investigators.

The p value and the theory of hypothesis testing are useful tools that help doctors conduct research. They are helpful for planning an experiment, interpreting the results observed, and reporting findings to peers. However, it is paramount researchers understand precisely what these tools mean and do not mean so that eventually they will not be blinded by the irrelevance of some statistical value in front of important medical reasoning.

Table 3. Implications for research

Step	Implication
Hypothesis giving rise to the research	The hypothesis tested should be relevant as determined by previous experiments, logical biologic or mechanical effect, etc
Planning	α, power, and sample size should be determined a priori.
Design and conduction	Study design should limit the biases so that differences observed may be attributable to the treatment or characteristic under scrutiny
Report	Methods should be detailed sufficiently so that an informed investigator may reproduce the research; discussion should report internal and external validity limits of the study
Confrontation	Study results should be confronted with previous and future results before the hypothesis tested is accepted or rejected

References

1. Bailey CS, Fisher CG, Dvorak MF. Type II error in the spine surgical literature. *Spine (Phila Pa 1976).* 2004;29:1146–1149.
2. Biau DJ, Kerneis S, Porcher R. Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clin Orthop Relat Res.* 2008;466:2282–2288.
3. Fisher RA. *Statistical Methods for Research Workers.* Edinburgh, UK: Oliver and Boyd; 1925.
4. Fisher RA. The arrangement of field experiments. *J Ministry of Agriculture Great Britain.* 1926;33:503–513.
5. Fisher RA. *Statistical Methods for Research Workers.* Ed 11 (rev). Edinburgh, UK: Oliver and Boyd; 1950.
6. Fisher RA. *Statistical Methods and Scientific Inference.* Ed 2 (rev). Edinburgh, UK: Oliver and Boyd; 1959.
7. Freedman KB, Back S, Bernstein J. Sample size and statistical power of randomised, controlled trials in orthopaedics. *J Bone Joint Surg Br.* 2001;83:397–402.
8. Garcia-Cimbrelo E, Diez-Vazquez V, Madero R, Munuera L. Progression of radiolucent lines adjacent to the acetabular component and factors influencing migration after Charnley low-friction total hip arthroplasty. *J Bone Joint Surg Am.* 1997;79:1373–1380.
9. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol.* 2008;45:135–140.
10. Goodman SN. Toward evidence-based medical statistics. 1: The p value fallacy. *Ann Intern Med.* 1999;130:995–1004.
11. Hodgkinson JP, Shelley P, Wroblewski BM. The correlation between the roentgenographic appearance and operative findings at the bone-cement junction of the socket in Charnley low friction arthroplasties. *Clin Orthop Relat Res.* 1988;228:105–109.
12. Hopkins PN, Williams RR. A survey of 246 suggested coronary risk factors. *Atherosclerosis.* 1981;40:1–52.
13. Hubbard R, Bayarri MJ. P values are not error probabilities. Available at: http://www.uv.es/sestio/TechRep/tr14-03.pdf. Accessed January 13, 2009.
14. Kobayashi S, Eftekhar NS, Terayama K, Iorio R. Risk factors affecting radiological failure of the socket in primary Charnley low friction arthroplasty: a 10- to 20-year followup study. *Clin Orthop Relat Res.* 1994;306:84–96.
15. Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc Lond A.* 1933;231:289–337.
16. Onsten I, Akesson K, Obrant KJ. Micromotion of the acetabular component and periacetabular bone morphology. *Clin Orthop Relat Res.* 1995;310:103–110.
17. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine.* 1900;5:157–175.
18. Schmalzried TP, Kwong LM, Jasty M, Sedlacek RC, Haire TC, O'Connor DO, Bragdon CR, Kabo JM, Malcolm AJ, Harris WH. The mechanism of loosening of cemented acetabular components in total hip arthroplasty: analysis of specimens retrieved at autopsy. *Clin Orthop Relat Res.* 1992;274:60–78.
19. Scott IA. Evaluating cardiovascular risk assessment for asymptomatic people. *BMJ.* 2009;338:a2844.
20. Sterne JA, Davey Smith G. Sifting the evidence what's wrong with significance tests? *BMJ.* 2001;322:226–231.